**Proc. Second International Conference on Speech and Language Processing, Vol. 2**
Banff, Alberta, Canada, October 12th.-16th. 1992, pp 1219-1222

# UNRESTRICTED TEXT-TO-SPEECH REVISITED: RHYTHM AND INTONATION

David R. Hill, Craig-Richard Schock and Leonard Manzara

Computer-Human Systems Lab
Dept. of Computer Science, U of Calgary,
Calgary, Alberta Canada T2N 1N4
email: hill@cpsc.ucalgary.ca

## ABSTRACT

A new speech-synthesis-by-rules system has been developed, at the University of Calgary, in an object-oriented programming environment, on the NeXT computer. This paper outlines the models used to create rhythm and intonation for the synthesised speech. A companion paper (DEGAS: a system for rule-based diphone speech synthesis) outlines the framework for segment specification [*Editor's note 2006*: DEGAS and the NeXT have been superceded by a GPL system that runs on a Macintosh under OS/X; see *http://savannah.gnu.org/projects/gnuspeech*].

The rhythm model is based on data obtained from real speech and represents a continuation of earlier work. The difficulties in reconciling the structure used for synthesis with the structure assumed for traditional segmental analysis are outlined. The intonation model is based on the descriptive framework developed by M.A.K. Halliday, as used for teaching non-native speakers to produce reasonable intonation for spoken English. The encouraging results from preliminary subjective testing of the intonation model, with utterances synthesised using the system, are described. On the basis of these tests, together with informal evaluation, we conclude that the speech produced represents a significant improvement in naturalness and intelligibility.

The research provided the basis for a new commercial text to-speech system that was marketed by *Trillium Sound Research Inc*—a successful technology transfer exercise. [*see note above*]

## 1 INTRODUCTION

### 1.1 Background

Unlike speech synthesised from compressed versions of real speech, speech synthesised by rules still leaves much to be desired in terms of naturalness, intelligibility and flexibility. Elaborate schemes have been devised for concatenating segments, or interpolating from one set of phonetic targets to another, using a variety of synthesis models, but fundamental problems remain. The realisation of arbitrary phonemes, in context, requires access to very detailed acoustic specifications based on phonetic context, speaking rate, and dialect, as well as the identity of the sound to be produced. Few systems successfully provide the detailed framework needed to capture, represent and use such information. The commonest response to the problem is to use prerecorded diphone segments from real speech, where the necessary detail is implicit in the underlying recordings.

Our system uses diphone elements synthesised by rules [1]. One of the earliest successful systems of this type was the system developed by Dixon and Maxey [2]. Witten [3] and Hill [4] provide reasonable surveys of the background to this work.

### 1.2 Rhythm and Intonation

*Segmental boundaries* Even if the segmental level of speech synthesis can be synthesised to a high standard[1], rhythm and intonation[2] are still difficult to specify, for an arbitrary utterance.

Rhythm is made up of segment durations and stressing features (perceived as prominence or salience), but segment durations imply some kind of boundaries. The nature and existence of such boundaries is contentious, and provides one of the sources of problems in our own work. In discussing the boundary assignment problem when developing the MiTalk system [5], after explaining where the boundaries were set for duration measurement, the authors say:

> These definitions lead to a convenient and largely reproducible measurement procedure, but the physiological and perceptual validity of these boundaries has not been established.

Clearly, the methodology and criteria for placing such boundaries is open to question, though there is often close agreement between analyses produced by different workers, if care is taken [6].

*Interaction between rhythm and intonation* Rhythm and intonation cannot be entirely separated. Pitch changes may lend prominence, providing a beat and thereby affecting rhythm, while rhythmic variation interacts with the detailed placement of the pitch features that make up intonation contours. Any intonation contour must be tied to the syllabic and rhythmic structure of the utterance. To complicate matters, varieties of microintonation—changes in pitch due to segmental performance constraints rather than intonational intent—must be reproduced, independently of the overall intonation and rhythmic structure. Microintonation is a cue to segmental identity, but good data are lacking. Finally, the suprasegmental resources of rhythm and intonation act in concert to augment and clarify the meaning of utterances. Conflicting cues can seriously degrade intelligibility and perceived speech quality.

Improving rhythm and/or intonation offers avenues by which we can improve both the quality (including naturalness) and the intelligibility of synthetic speech. Regrettably, there seems to be a dearth of detailed, experimentally derived published information on which to base algorithms for good rhythm and intonation when synthesising speech by rules. Excellent data comes from the IPO [7, 81. Papers co-authored by one of the present authors are also relevant [6,9, 10].

*Rhythm* The rhythmic character of speech may be specified as the occurrence of "beats", and other features, at intervals that vary. The overall framework frequently adopted for handling rhythm is reminiscent of measures in musical notation. Each rhythmic unit begins with a beat, or stressed syllable. Some syllables, those at "information points" in the phrase or sentence, bear what may be called *tonic stress*, and are therefore particularly prominent. The time intervals between the beats (or other rhythmic features) obviously depends directly on the durations of the segments included in the interval. This is generally what people mean when they talk about rhythm.

The perception of beats themselves often depends partly on a duration distinction—longer means more prominent. We have also found the rhythm internal to words is important for intelligibility. For all these reasons, the primary determinant of rhythmic structure is the allocation of duration to segments.

[1] And there are serious problems at this level, regardless of whether synthetic or real speech is involved.

[2] Rhythm and intonation will be referred to collectively as "suprasegmental features", "suprasegmentals" etc., although these term are really more general.

A contentious issue in designing and fleshing out the rhythmic framework is the question of isochrony. English is a *stress-timed language*—perceived rhythm depends on the pattern of beats rather than syllables. The British linguistic tradition, following scholars like Jones, Abercrombie, Ladefoged and Halliday, asserts that the beats tend to occur at regular (isochronous) intervals, and that the segments vary in length to achieve this end. The American tradition asserts that the effect, if it exists at all, is mainly a perceptual effect, and cite huge variability in interval lengths (ratios of 6 or 7 to 1 between longest and shortest inter-beat intervals) to support their argument. We have evidence to support both views, oddly enough. The range of interval lengths is real, but there is still some residual length variation in response the to number of segments intervening between beats, accounting for some 10% of the variance in segment durations (see Section 3).

The notation used in Halliday's book [11], although designed primarily to show intonational structure, inevitably has to provide a framework for the rhythm, for reasons already stated.

*Intonation* Intonation comprises the pattern of variation in pitch frequency over many segments, and tends to relate to grammatical units, as well as the message the speaker wishes to convey. Information points and phrase boundaries are often marked by particular features of the intonation pattern, such as rises and falls, the details of which depend on their function. Halliday [11] has specified a simple categorisation for the basic patterns found in British English. The IPO workers have a more complex grammar to perform a similar function [7, 8, 12].

For unrestricted text-to-speech, the choice of intonation contours required for stretches of speech is, in general, not known. Intonation, like rhythm, can dramatically affect meaning both denotative and connotative, as well as other attributes, such as emotion. Ultimately, correct intonation requires understanding the speech and context in a way that computers are, as yet, unable to accomplish. In addition, segmental effects on pitch variation (microintonation) should be incorporated automatically. While constructing synthetic intonation contours algorithmically appears to be a logical step, it is important to remember that intelligibility can be diminished, and meaning changed, if unacceptable contours are used. It is with this in mind that the quality of intonation contours becomes an issue of importance in any text-to-speech system. An experiment, described below, provides a rationale for the approach we have adopted to generating synthetic intonation.

### 1.3 Segmental analysis

Phoneticians traditionally segment speech into successive *phones*. One goal of the field linguist is to discover these phones for each language to be investigated and classify them into categories called *phonemes*. Sounds belong in the same category, roughly speaking, if they do not distinguish two words in the language concerned. The engineer trying to synthesise or recognise speech has a tendency to assume that these phonemes categories are identifiable entities in their own right. They are not. Sounds (phones) in the same phoneme category may differ just as upper and lower case "A/a" differ. Consider the "l" sounds in "blue" and "clear". A further incorrect assumption is that phones follow each other in speech like the bricks in a wall. They do not. They interact, overlap, and modify each other. A phonological view approximates this reality more closely. A clear account of many of the issues appears in O'Connor's excellent book [13], recently reprinted.

In order to obtain the data needed to construct parameter tracks at the segmental level[3], or to construct a rhythm model and use it, segment boundaries must be assumed, and inserted in the original data, whether they exist or not. The boundaries are then reproduced in some form when synthesising speech. This causes serious problems when attempting algorithmic specifications suited to computer implementation. What is particularly frustrating is that traditional phonetic analysis (phonetic, not phonological) is not even self-consistent. It was designed for a different purpose, of course, and has proved its value for that.

We are still trying to resolve the problem, as described below. We prefer to talk about *postures*, instead of phones, since these have an innate articulatory reality, and lead naturally to the required interaction, overlap and cross-modification, when used as the basis for constructing synthetic speech. We also use an *event based* approach to synthesis, which is more of a phonological than a phonetic view of the speech. Postures are defined by the target values of acoustic parameter specifications, one set for each posture. This reflects the idea that, when an articutatory posture is achieved, the vocal tract will be in a certain configuration, and will generate acoustic data directly related to this configuration. As the articulators move to the next posture (a *vocal gesture*), dynamic changes are generated that may be mimicked by appropriate interpolation rules (we call them *transition profiles*) from one set of targets to the next. "Event -based" simply means that we are not particularly constrained by notional "boundaries", at least in principle. Noise bursts may begin and end at arbitrary times; formant transitions likewise; and so on.

Unfortunately, much acoustic-phonetic data is cast in segmental terms—even our own. Such data inherits the difficulties and inconsistencies of the analysis. It is not suited to our revised view of speech structure, and compromises our ability to create a less constrained synthesis-by-rules system—if only because of the monumental difficulty of collecting all the required data ourselves. We compromise in order to use existing data, but are looking for ways of circumventing the problems.

## 2 PROBLEMS SYNTHESISING (DIPHONE) SEGMENTS

When performing segmental analysis of spectrographic data, the phonetician places boundaries between successive sound segments, based on visual criteria. Computer analysis may mimic this procedure using related spectral change criteria, and the like. The process sounds simple and repeatable. However, dubious assumptions are built into the method; and convention rather than real acoustic evidence may dictate certain decisions. For example, *contoids* (roughly consonants) are distinguished by the degree of constriction in the vocal tract. Thus, in segmenting contoids, the boundaries are placed so as to include only the most extreme deviation of the formant parameters. The formant transitions will be assigned to any neighbouring vocoids (vowel-like sounds), if present. When contoids abut, a somewhat different criterion must be used so that the duration of a given contoid will reflect some variance due only to a change in segmentation criteria. Conventional wisdom declares that there are long vowels and short vowels. Experience shows that there are at least three categories of vowel duration, within the same phonetic vowel quality, without taking account of the notions of vowel reduction, open transitions, and the like.

Such considerations probably mean that synthesis based on traditional phonetic analysis can only be a very rough approximation. The obvious solution is to collect data on all postures, in all contexts. Strictly speaking, analysis, like synthesis, needs to be based on no less than diphone units—possibly larger units. The perception and recognition of spoken words may be much more closely akin to the perception and recognition of Chinese characters, than words made up from the Roman alphabet. In order to achieve the required numbers for statistical significance in the multitude of potential categories, the labour involved in

[3] The reader should now see that the term "segmental level" is misbegotten, since "segmental" implies divisions between segments. However, the term is traditional, and we shall to use it, with this caveat.

formal analysis and abstraction is beyond the resources of all but the most well-endowed organisations. Even these may be unwilling to commit the necessary resources. There has been a consequent surge in the use of ready-made data derived from real speech and tailored to the constraints of synthesis using modern acoustic analysis and editing tools. While possibly satisfying the need for immediate improvements in speech synthesis quality, the approach tells us little about the formal structure of speech as needed for more flexible, universal synthesis, or for speech recognition. In this sense, such approaches are *ad hoc* and somewhat limited.

## 3 SYNTHESISING RHYTHM

Our rhythm model is based on segment duration data obtained from a detailed segmental analysis of the published tape recordings for Study Units 30 and 39 in Halliday's book [11]. The book is designed to teach acceptable British English intonation to non-native speakers. The underlying data and analysis for our model were presented in [10]. A full length revised version, based on recent experience, is in preparation.

Our model may be characterised by the sources contributing to the variance in duration of segments. Currently, these include: the segment identity; tonic placement and finality of the rhythmic unit (collectively designated *marked* as opposed to other rhythmic units which are *unmarked*); and the isochrony effect. Together they account for roughly 75% of the observed variance in segment duration. Correctly specifying all of these ultimately requires understanding of the text, or intent in speaking. As our system currently does not even parse the text, defaults must be provided to indicate marked rhythmic units, in the absence of supplementary information provided by the user. Our system does make some provision for the user to provide this information, if desired.

Dictionary look-up specifies the segment identities, and identifies syllables capable of taking stress—thereby automatically dividing the speech up into rhythmic units. Segment durations are then assigned, based on the three criteria just mentioned, the isochrony effect taking the form of a linear regression, whilst the other factors are covered by table look-up.

Our biggest problems are the discrepancies between the analysis framework and the synthesis framework, and the lack of enough detailed formal data.

## 4 AN EXPERIMENT ON INTONATION

### 4.1 Purpose and methodology

The experiment was designed to evaluate the relative naturalness of synthetic intonation contours generated by two sets of rules, prior to using the rules in our text-to-speech system. The points at issue were: (a) whether synthetic intonation contours are *acceptable* replacements for natural contours; and (b) which of the available alternatives is most attractive. The result surprised us.

Ten sentences were chosen from Study Unit 30 in Halliday's book [11]. All ten sentences were synthesised by rules, using our text-to-speech system, but invoking raw input mode (a method of providing direct phonetic specification as input) in order to obtain the closest phonetic representation of the tape recordings supplied with the book, which we had previously analysed. Three versions of each sentence were synthesised. One used a copy of the original intonation contour to control the pitch. The second used a synthetic contour provided by a simplified version of the IPO rules for synthesising British English intonation [7]. The third used a synthetic contour provided by a simplified version of Halliday's intonation specification, based on an algorithm implemented by Witten [3, pp 201-207], but extended somewhat to include contours missing from the original Witten implementation.

Although some of the resulting synthetic contours had obvious timing discrepancies between features of the intonation contours and the speech segments, no corrections were applied, as we needed to evaluate the algorithms exactly as given.

A total of fifteen subjects heard each version of each sentence three times, in randomised order. They were shown the set of sentences in text form prior to starting the experimen, and a display of each sentence was also shown before it was spoken. Each subject listened to 90 sentences. Subjects ranked the sentences on a six point scale:

1 Very unnatural
2 Quite unnatural
3 Somewhat unnatural
4 Somewhat natural
5 Reasonably natural
6 Very natural.

A graphical user interface for the experiment, running on the same NeXT computer as the synthesis, allowed subjects to proceed at their own pace. Feedback on the number of sentences already heard was provided, along with an analog display of percentage complete. However, subjects were only allowed to hear each sentence once. The results were recorded automatically.

### 4.2 Results and analysis

The results comprised average ratings on the six point scale for each sentence, and each contour type—a total of 30 ratings. Using Edwards method of successive intervals [14], the data were transformed to a new parametric acceptability scale to which analysis of variance (ANOVA) could be applied. Figure 1 shows the mean scores by utterance and intonation contour type.

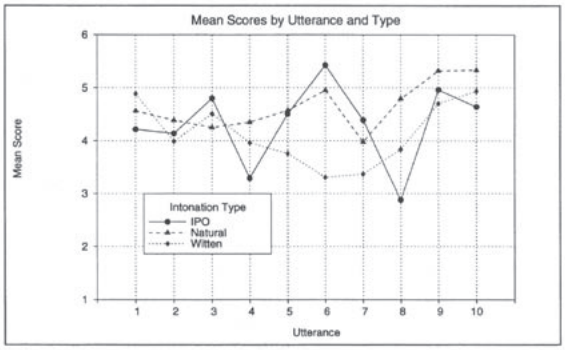

*Figure 1: Mean naturalness scores by utterance and type*

The ANOVA showed there was a significant interaction between intonation type and utterance. That is, some utterances scored higher or lower than might be expected for their intonation type, as clearly illustrated in Figure 1, where the main culprit is seen to be the IPO-type contour. In fact the variability in perceived naturalness was extreme for this contour type. Sentence 6 was judged more natural than nearly all other utterances, regardless of contour type, whilst sentences 4 and 8 were worse than all other utterances.

Because such variation is not too surprising, given the somewhat arbitrary nature of the contour assignment, and since the main effects still produced significant differences, we considered continued analysis to be appropriate. However, the extreme variability of the IPO-type results is relevant when it comes to selecting an approach to synthetic contour generation, as noted below.

The main treatments (sentence and contour type) also produced statistically significant differences in judgements of naturalness. We were particularly interested in the latter. We performed a *post hoc* Tukey HSD test of the mean scores by type, to find out which differences between means led to the ANOVA finding of a statistically significant effect from contour type. There were statistically significant differences between the mean scores for natural contour type sentences and both the IPO and Halliday types, but not between the IPO and Halliday types. It is also interesting that the mean scores for all three types fell in the range between “Somewhat natural” and “Reasonably natural”.

### 4.3 Discussion and interpretation

*Summary* Five main points arise from the results obtained:

(a) Both our study, and other previous studies [7, 8, 12] report a difference in the acceptability ratings of the natural and synthetic intonation schemes;

(b) The differences in mean score between the IPO and Halliday contours are neither statistically nor practically significant;

(c) All contour types had mean scores in the range somewhat natural to reasonably natural;

(d) There was great variability in the perceived naturalness of the contours produced by the simplified IPO intonation method we used; and

(e) Even the simplified IPO method was more complicated to apply than our extension of the Witten algorithm for the Halliday contours.

*Implications for practitioners* Compared to the extended Witten implementation of Halliday’s system, the IPO approach is much more complete system, backed by considerable in-depth perceptual research. However, the IPO system achieved only slightly better mean scores whilst being considerably more complicated to implement, and showing extreme variability. The apparent perceptual advantage over the Witten system was not statistically significant, and the Witten system produced more uniform naturalness scores, as may be seen in Figure 1.

Consistency and simplicity are seen as virtues for a text -to speech system, all other things being effectively equal.

Possible confounding factors in our experiment included the fact that all contours were representations of British English contours, but were judged mainly by Canadian English speakers. Thus the overall naturalness judgements may have been depressed. This is not considered a problem, since the bias, if it existed, presumably applied equally to all utterances. Also, for reasons of resource limitations, the subjects were not given much practice, apart from a small training set prior to starting the experiment, and several said their first ten (or so) ratings were uncertain. The random re-ordering of presentation sequence for each subject should have eliminated any problem from this source. We do not know if the use of strictly synthetic segmental speech structure affected our results and, if so, how. Also, presenting pairs of utterances bearing randomly selected contour types, and forcing listeners to say which sounded more natural, could have been more sensitive to differences. Finally, a more complete implementation of the IPO system may have produced more consistent and higher scores for that method.

Because the extended Witten approach to intonation synthesis is simplest and most consistent, and does not generate significantly less natural contours than the alternative IPO scheme, we feel it is currently the method of choice for practitioners. We have dealt with the timing discrepancies by tying the intonation contour features, such as start of tonic fall, to the nuclear vowel of the stressed syllable of the relevant rhythmic unit. Previously it was tied to the rhythmic unit boundary. Small changes like this make a surprising difference, and we have already improved the synthetic intonation algorithm compared to the earlier version used for the experiment. Further research is planned.

## 5 CONCLUSIONS

The rhythm model used for synthesis is based on real speech, and accounts for approximately 75% of the natural rhythmic structure. None of the subjects commented on the rhythm, when the model was used to generate utterances for testing approaches to intonation and, though it possibly had some effect on the results, we have no evidence that it did. Further experimentation is needed. The biggest obstacle to improving the segmental structure, and hence the rhythm, is collecting and formalising a vast amount of new data, based on a more realistic segmental framework.

The extended Witten approach to modelling Halliday’s intonation system for spoken English was chosen, in preference to the more complex IPO approach, because both were acceptably natural, but the Witten was simpler and more consistent in naturalness. The most important difficulty in this part of the research was tying the features marking each chosen intonation contour to the segmental structure, choosing an appropriate contour, and dealing with other than default placement of the tonic rhythmic unit. Progress in the latter requires solution of one of the hard problems of Artificial Intelligence. However, the current system has already exceeded our expectations.

## 6 ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada under grant number OGP0005261. They also wish to thank Natascha Schuler for her participation in the design and running of the intonation experiment. Finally, they would like to thank John Kendall for helping to provide the NeXT computer system without which this research would have been impractical.

## 7 REFERENCES

[1] L. Manzara and D.R. Hill, “DEGAS: a System for Rule Based Diphone Speech Synthesis.” *ICSLP 92*, Banff, Oct. 12-16 1992. [*Editor’s note 2006*: DEGAS and the NeXT system have been superceded by a GPL system that runs on a Macintosh under OS/X; see *http://savannah.gnu.org/projects/gnuspeech*]

[2] N.R. Dixon and H.D. Maxey, “Terminal Analog Synthesis of Continuous Speech Using the Diphone Method of Segment Assembly.” *IEEE Trans. on Audio and Electro-Acoustics*, vol. AU-16, pp. 40-50, 1968.

[3] I.H. Witten, *Principles of Computer Speech*. London: Academic Press, 1982.

[4] D.R. Hill, “Spoken Language Generation and Understanding by Machine: a Problems and Applications Oriented Overview. *Spoken Language Generation and Understanding* (J.C. Simon, ed.), Dordrecht: D. Reidel, pp.3-38, 1980.

[5] J.Allen, M. Sharon Hunnicutt and Dennis Klatt, *From text to speech: The MITalk system*. Cambridge, UK: Cambridge U. Press, 1987.

[6] D.R. Hill, W. Jassem, and l.H. Witten, “A Statistical Approach to the Problem of lsochrony in Spoken British English.” *Current Issues in Linguistic Theory* (Amsterdam: John Benjamins BV), vol. 9, pp. 285-294, 1979 .

[7] N. Willems, R. Collier, and J. ’t Hart, “A Synthesis Scheme for British English intonation.” *J. Acoustical Soc. America*, vol. 84, pp. 1250-1261, 1988.

[8] J. ’t Hart, R. Collier and A. Cohen, *A Perceptual Study of Intonation*, Cambridge, England: Cambridge U. Press 1990,

[9] D.R. Hill and NA, Reid, “An Experiment on the Perception of Intonational Features.” *Int. J. Man-Machine Studies*, vol. 9, pp. 337-347, 1977.

[10] D.R. Hill, I.H. Witten and W. Jassem, “Some Results from a Preliminary Study of British English Speech Rhythm.” *94th. Meeting of the Acoustical Society of America*, Miami, Dec. 1977.

[11] M.A.K. Halliday, *A Course in Spoken English: Intonation*. Oxford: Oxford U. Press 1970.

[12] J.R. de Pijper, *Modelling British Intonation*. Rhode Island: Cinnaminson, Dordrecht: Foris, 1983.

[13] J.D. O’Connor, *Phonetics*. Harmondsworth, England: Penguin Books, 32Opp. 1973.

[14] A.L. Edwards, *Techniques of Attitude Scale*. Appleton Century-Crofts, 1957.